Assignment - 2

1. Theoretical Part.

1.1 a) Given the tanh Activation function. to revise backpropagation Algorithm.

$$\tanh(x) = \frac{e^{x} + e^{-x}}{e^{x} + e^{-x}}$$

Here $\sigma(x) = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

$$\frac{\partial \sigma(x)}{\partial x} = \frac{(e^{x}+e^{-x})(e^{x}+e^{-x}) - (e^{x}-e^{-x})(e^{x}-e^{-x})}{(e^{x}+e^{-x})^2}$$

$$1 - \left[\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\right]^2 = 1 - (\tanh x)^2$$

$$= 1 - (\sigma(x))^2$$

Error for Example d is:

$$E_d(w) = \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2$$

$$\Delta w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}}$$

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial net_j} \times \frac{\partial net_j}{\partial w_{ji}}$$

Case 1: j is an output unit.

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}$$

$$E_d(w) = \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2$$

$$\frac{\partial E_d(w)}{\partial o_j} = \frac{\partial}{\partial o_j}\left[\tfrac{1}{2}(t_j - o_j)^2\right] = -(t_j - o_j)$$

$$\frac{\partial o_j}{\partial net_j} = [1 - o_j]^2$$

$$\frac{\partial E_d}{\partial net_j} = -(t_j - o_j) \times (1 - o_j^2) = -\delta_j$$

Let $\delta_j = (t_j - o_j) \times (1 - o_j^2)$

$$\Delta w_{ji} = -\eta \times \frac{\partial E_d}{\partial w_{ji}} = \eta (t_j - o_j)(1 - o_j^2)$$

Case 2: j is a hidden unit

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in Downstream(j)} \frac{\partial E_d}{\partial net_k} \times \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial net_j} = \sum_{k \in Downstream(j)} -\delta_k \times \frac{\partial net_k}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}$$

$$= \sum_{k \in Downstream(j)} -\delta_k w_{kj} \times (1 - o_j^2).$$

$$\delta_j = (1 - o_j)^2 \sum_{k \in Downstream} \delta_k w_{kj}$$

Backpropagation Summary: For tanh Activation

$$\Delta w_{ji} = \eta \delta_j x_{ji}$$

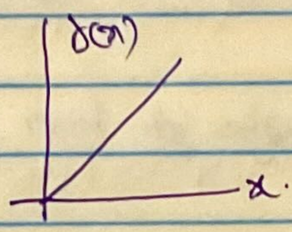
Case 1: $\delta_j = (t_j - o_j)(1 - o_j)^2$

Case 2: $\delta_j = (1 - o_j)^2 \sum_{k \in Downstream} \delta_k w_{kj}$

b) Given the Relu Activation function to revise the backpropagation Algorithm.

Case 1: When j is the output layer

$$Relu(x) = \begin{cases} \max(0, x) \\ 0 \end{cases}$$



$$g'(x) = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

$$\frac{\partial E_d}{\partial net_j} = \frac{\partial E_d}{\partial o_j} \times \frac{\partial o_j}{\partial net_j} \qquad \frac{\partial o_j}{\partial net_j} = \begin{cases} 0 & x < 0 \\ 1 & x > 0 \end{cases}$$

$$E_d(\omega) = \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2$$

$$\frac{\partial E_d}{\partial o_j} = \frac{\partial}{\partial o_j}\left[\frac{1}{2}(t_j - o_j)^2\right] = -(t_j - o_j)$$

$$\Rightarrow \frac{\partial E_d}{\partial net_j} = \begin{cases} -(t_j - o_j) \times 0 & x < 0 \\ -(t_j - o_j) \times 1 & x > 0 \end{cases}$$

$$\Delta \omega_{ji} = \begin{cases} 0 & x < 0 \\ \eta(t_j - o_j) & x > 0 \end{cases}$$

Case 2: When j is the hidden unit

$$\frac{\partial E_d}{\partial net_j} = \sum_{k \in Downstream(j)} \frac{\partial E_d}{\partial net_k} \cdot \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial net_j}$$

$$= \sum_{k \in Downstream(j)} -\delta_k \frac{\partial net_k}{\partial o_j} \times \frac{\partial o_j}{\partial net_j}$$

$$= \sum_{k \in Downstream(j)} -\delta_k \times w_{kj} \times \begin{cases} 0 & x<0 \\ 1 & x>0. \end{cases}$$

$$\delta_j = \begin{cases} 0 & x<0 \\ \sum_{k \in Downstream} \delta_k w_{kj} & x>0 \end{cases}$$

Summary for Backpropagation Algorithm

i) For each output unit j

$$\delta_j \leftarrow \begin{cases} 0 & x<0 \\ (t_j - o_j) & x>0 \end{cases}$$

ii) For each hidden unit k.

$$\delta_k \leftarrow \begin{cases} 0 & x<0 \\ 1 \times \sum_{k \in Downstream} w_{ji} \delta_k. \end{cases}$$

iii) $w_{i,j} \leftarrow w_{ij} + \Delta w_{ij}$

$$\Delta w_{i,j} = \eta \delta_j x_{ij}$$

1.2. Given to derive gradient descent training rule for single neuron with output

$$o = w_0 + w_1(x_1 + x_1^2) + \cdots + w_n(x_n + x_n^2).$$

where $x_1, x_2 \ldots x_n$ are inputs

$w_1, w_2 \ldots w_n$ are weights correspondingly

To minimize the error we need to adjust the weights

$$w_i \leftarrow w_i + \Delta w_i$$

$$\Delta w_i = -\eta \nabla E(\vec{w})$$

$$= -\eta \frac{\partial E}{\partial w_i}$$

Training Error $E(\vec{w}) = \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$

where $D \rightarrow$ set of training examples

$t_d \rightarrow$ target o/p for training example d

$o_d \rightarrow$ o/p of the linear unit for training example d

$$\frac{\partial E}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_{d \in D} (t_d - o_d)^2$$

$$= \frac{1}{2} \sum_{d \in D} \frac{\partial}{\partial w_i} (t_d - o_d)^2$$

$$= 1/2 \sum_{d \in D} 2(t_d - o_d) \frac{\partial}{\partial w_i} (t_d - o_d)$$

Given output $o = w_0 + w_1(x_1 + x_2^2) + \cdots w_n(x_n + x_n^2)$

vectorized o/p will be $o = \vec{w}(\overrightarrow{x_d + x_d^2})$

$$= \sum_{d \in D} (t_d - o_d) \frac{\partial}{\partial w_i} (t_d - \vec{w}(\overrightarrow{x_d + x_d^2}))$$

$$\frac{\partial E}{\partial w_i} = \sum_{d \in D} (t_d - o_d)(-x_{id} - x_{id}^2).$$

So, the final result for the weight update for the gradient descent will be

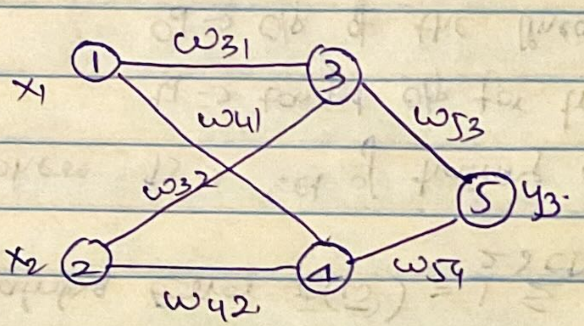
$$\Delta\vec{w} = -\eta\left[\sum_{d\in D}(t_d-o_d)(-x_{id}-x_{id}^2)\right]$$

$$= \eta\sum_{d\in D}(t_d-o_d)(x_{id}+x_{id}^2).$$

x_{id} denotes the single input component x_i for training example d.

1.3

a)



$$y_3 = h(w_{31}x_1 + w_{32}x_2)$$
$$y_4 = h(w_{41}x_1 + w_{42}x_2)$$
$$y_5 = h(w_{53}y_3 + w_{54}y_4)$$
$$y_5 = h\left[w_{53}h(w_{31}x_1+w_{32}x_2) + w_{54}h(w_{41}x_1+w_{42}x_2)\right]$$

b)

$$X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad w^{(1)} = \begin{pmatrix} w_{3,1} & w_{3,2} \\ w_{4,1} & w_{4,2} \end{pmatrix}$$

$$w^{(2)} = (w_{5,3} \;\; w_{5,4})$$

output of hidden layer $= h(w^{(1)}x)$

output at output layer $= h(w^{(2)} \cdot h(w^{(1)} \cdot x))$.

c) Relationship between $hs(x)$ and $ht(x)$

$$ht(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$= \frac{e^{x} + e^{-x} - 2e^{-x}}{e^{x} + e^{-x}} = 1 - \frac{2}{e^{2x} + 1}$$

$$ht(x) = 1 - \frac{2}{e^{2x} + 1} = 1 - 2hs(-2x).$$

$$\because hs(-x) = \frac{1}{1+e^{x}} = 1 - \frac{1}{1+e^{-x}} = 1 - hs(x)$$

$$ht(x) = 1 - 2(1 - hs(2x)).$$
$$= 1 - 2 + 2hs(2x)$$
$$ht(x) = 2hs(2x) - 1$$

Here we can see Sigmoid and tanh Activation functions have a linear relationship.
So Two Activation functions can generate the same function.